Posterior Mean Super-resolution with a Causal Gaussian Markov Random Field Prior

Takayuki Katsuki, Akira Torii, and Masato Inoue

***Abstract*—We propose a Bayesian image super-resolution (SR) method with a causal Gaussian Markov random field (MRF) prior. SR is a technique to estimate a spatially high-resolution image from given multiple low-resolution images. An MRF model with the line process supplies a preferable prior for natural images with edges. We improve the existing image transformation model, the compound MRF model, and its hyperparameter prior model. We also derive the optimal estimator – not the joint maximum a posteriori (MAP) or marginalized maximum likelihood (ML), but the posterior mean (PM) – from the objective function of the L2-norm (mean square error) -based peak signal-to-noise ratio (PSNR). Point estimates such as MAP and ML are generally not stable in ill-posed high-dimensional problems because of overfitting, while PM is a stable estimator because all the parameters in the model are evaluated as distributions. The estimator is numerically determined by using variational Bayes. Variational Bayes is a widely used method that approximately determines a complicated posterior distribution, but it is generally hard to use because it needs the conjugate prior. We solve this problem with simple Taylor approximations. Experimental results have shown that the proposed method is more accurate or comparable to existing methods.**

***Index Terms*—super-resolution, Bayesian inference, Markov random field prior, line process, posterior mean, variational Bayes, Taylor approximation.**

I. Introduction

Super-resolution (SR) is an information processing technique that makes it possible to infer a spatially high-resolution (HR) image of a scene from corresponding multiple low-resolution (LR) images that are affected by warping, blurring, and noise. SR can be applied to a variety of images; e.g., still images extracted from several sequential video frames. SR needs the registration of LR images in addition to the image restoration of the registered LR images. Since the earliest work by Tsai and Huang [1], SR has been achieved using various methods [2]–[10] and good overviews of these methods are given in [11]–[16]. Generally, SR is an ill-posed inverse problem because inverting the blur process without amplifying the effect of the noise is difficult [13]. In other words, the degrees of freedom of the HR image and pixel-wise observation noise are always higher than the dimensionality of the observed LR images, so complete determination of an HR image is impossible. Therefore, the HR image is frequently inferred as the most preferable image within the framework of the probabilistic information processing, and we handle SR using this framework in this paper. The probabilistic information processing has three key features: 1) model, 2) objective function, and 3) optimization method. In the SR problem, the model includes the observation model and the prior model. The observation model consists of warping, blurring, downsampling, and noise models. The prior model, necessary for the Bayesian framework, mainly consists of an HR image prior, and sometimes includes both the hyperparameter prior for the HR image prior and the registration prior. The objective function evaluates how good or bad an estimator is. The estimator usually represents the inferred HR image, and sometimes includes auxiliary parameters; e.g., the registration parameters and edge information. The optimization method numerically maximizes/minimizes the objective function and determines the estimator. An optimization method is not necessary for simple problems in which an analytical exact solution can be obtained. In the probabilistic information processing, SR can be categorized according to these three key features.

To deal with warping, blurring, and downsampling, a linear transformation model is frequently used [3], [6], [8], [10]. Warping is usually limited with planar rotation and parallel translation. Blurring is defined by using a point spread function (PSF); a square or Gaussian type PSF is common. Downsampling denotes sampling from an HR image to construct an LR image. Downsampling sometimes includes anti-aliasing. Since these three transformations are linear, they can be combined into a single transformation matrix. As for the noise model, pixel-independent additive white Gaussian noise (AWGN) is usually employed.

The Bayesian framework, especially the HR image prior, is quite useful for SR. The HR image prior provides appropriate smoothness between neighboring pixel luminances. A common type of HR image prior imposes an L2-norm penalty on differences between horizontally and vertically adjacent pixel luminances (the first derivative). The L1-norm of the first derivative is sometimes used, and it has the advantage of robust inference against outliers. The total variation (TV) prior [10] employs the L1-norm of the gradient vector. The Huber prior [5] is a mixture prior of L1- and L2-norms. The SAR model [2], [9], [17] employs the response of a two-dimensional Laplacian filter (the second derivative). The Gaussian process prior [3] has neighboring pixels spread according to a Gaussian distribution. Besides the degree of smoothness between neighboring pixels, information regarding the discontinuity, or equivalently, the edges or line process, is also useful for inference. A common type of prior implementing edges is the compound Markov random field (MRF) prior that was introduced by Geman & Geman [18] and is widely used [4], [6], [8]. With respect to the compound MRF

Takayuki Katsuki, Akira Torii, and Masato Inoue are with the Department of Electrical Engineering and Bioscience, Graduate School of Advanced Science and Engineering, Waseda University, 3–4–1, Okubo, Shinjuku, Tokyo 1698555, Japan. E-mail: (see http://www.eb.waseda.ac.jp/m_inoue/).

[19], [20] prior, the normalizing constant, or equivalently, the partition function, is usually difficult to calculate because it has an exponential calculation cost with respect to the dimensionality of the line process. Recently, Kanemura *et al.* [6], [8] confusingly introduced a "causal" type of Gaussian MRF prior whose calculation cost is polynomial. We try to improve this prior in this paper.

The SR estimator should be derived from an objective function. As the objective function, a posterior distribution has been widely employed. Since the posterior distribution usually includes both the HR image and registration parameters, the joint maximum a posteriori (MAP) solution [2] is a suitable estimator for this objective function. Other than the joint MAP, the use of the marginalized maximum likelihood (ML) [3], [6] or marginalized MAP [5] has been proposed. Tipping *et al.* [3] and Kanemura *et al.* [6], [8] determine the registration parameters by using ML inference, where the HR image is marginalized out, and determine the HR image by using MAP inference. Pickup *et al.* [5] determines the HR image by using MAP inference, wherein the registration uncertainties are marginalized out, and assumes that the registration parameters are pre-registered by using standard registration techniques. Marginalized ML is also called type-II ML, evidence approximation, or empirical Bayes. Marginalized ML has no registration prior, unlike marginalized MAP. Pickup *et al.* [5] reported that marginalized MAP is superior to both joint MAP and marginalized ML. We evaluate the accuracy of SR methods in terms of the L2-norm (mean square error) - based peak signal-to-noise ratio (PSNR). Therefore, we think it is natural to employ PSNR as the objective function. For this objective function, posterior mean (PM) is a suitable estimator. The variational Bayes [21] approach [10] seems to approximately determine the PM of the HR image, although the authors assume some registration parameters are known and use point-estimate model parameters obtained by ML inference. To determine the exact PM of the HR image, all parameters other than the HR image should be marginalized out over the joint posterior distribution.

The type of optimization method to use is not as substantial a problem as the choice of model and objective function, but it is still important. Since almost all good estimators cannot be exactly determined because of difficult analytical integration or an exponential calculation cost, some approximation methods need to be introduced. Also, parameter tuning is necessary in many numerical optimization methods; e.g., of the initial value and the step-width settings in gradient methods. Specifically, in early work done on image restoration, an annealing method was used for the joint MAP solution [18], [22]. For marginalized ML and marginalized MAP solutions, the scaled conjugate gradients algorithm was used [3], [5]. In recent work, the variational expectation-maximization (EM) algorithm has been applied, which includes the gradient method in the M step [6], [8]. The variational Bayes approach has also been applied [10]. This method includes nested optimization of the majorization-minimization approach. This majorization-minimization approach seems to affect both the HR image prior and the estimator. Specifically, it modifies the TV prior to include a discontinuity parameter (called local spatial activity). In addition, this parameter is point-estimated when the HR image is inferred.

In this paper, we propose a new SR method that employs a "causal" Gaussian MRF prior and utilizes variational Bayes to calculate the optimal estimator, PM, with respect to the objective function of the L2-norm-based PSNR. This is a straightforward approach, but it was not proposed earlier possibly because an important limitation of variational Bayes is that a conjugate prior is needed. We solve this problem through simple Taylor approximations. In Section II, we define models, where we introduce a novel unified warping, blurring and downsampling model, an improved HR image prior, an improved hyperparameter prior, and a registration prior. In Section III, we employ PSNR as the objective function and derive the optimal estimator, PM, from this objective function. In Section IV, we determine the PM by using variational Bayes and Taylor approximations. In Section V, we evaluate the proposed method by comparing it with existing methods. We discuss the proposed method in Section VI and conclude in Section VII.

II. Model

A. Definitions

First, we define the gamma, Bernoulli, and Gaussian distributions used in this paper:

$$\begin{aligned}
\mathrm{Gamma}(x;a,b) &\equiv \frac{b^a}{\Gamma(a)}x^{a-1}\mathrm{e}^{-bx} \quad (x>0),\\
\mathrm{Bernoulli}(x;\mu) &\equiv \mu^x(1-\mu)^{1-x} \quad (x\in\{0,1\}),\\
\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &\equiv |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (\boldsymbol{x}\in\mathbb{R}^d),
\end{aligned}$$

Here, Γ is the gamma function, $|\bullet|$ denotes the determinant of a given matrix, superscript $\top$ denotes the transpose, $\mathbb{R}$ is the real number field, and d is the dimension of $\boldsymbol{x}$. The logistic function and Kullback-Leibler (KL) divergence from distributions $p(\boldsymbol{x})$ to $q(\boldsymbol{x})$ are respectively defined as

$$\begin{aligned}
\mathrm{logistic}(x) &\equiv \frac{1}{1+\mathrm{e}^{-x}},\\
D_{\mathrm{KL}}(p(\boldsymbol{x})\|q(\boldsymbol{x})) &\equiv \left\langle \ln\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right\rangle_{p(\boldsymbol{x})},
\end{aligned}$$

where the angle brackets $\langle\bullet\rangle_\circ$ denote the expectation of $\bullet$ with respect to a distribution $\circ$. Additionally, tr denotes the trace of a given matrix. diag denotes a diagonal matrix. $\boldsymbol{I}$ is an identity matrix of appropriate size. $\boldsymbol{0}$ is a zero vector or a zero matrix of appropriate size. All the vectors in this paper are column vectors. The $\|\bullet\|_2$ denotes the L2-norm of a given vector. At this point, these variables have absolutely nothing to do with the variables that appear later.

B. Observation Model

Our task is to estimate an HR grayscale image, $\boldsymbol{x}\in\mathbb{R}^{N_{\boldsymbol{x}}}$, from the observed multiple LR grayscale images, $\boldsymbol{Y}\equiv\{\boldsymbol{y}_l\}_{l=1}^L, \boldsymbol{y}_l\in\mathbb{R}^{N_{\boldsymbol{y}}}$. Images $\boldsymbol{y}_l$ and $\boldsymbol{x}$ are regarded as lexicographically stacked vectors. The number of pixels for each LR image, $N_{\boldsymbol{y}}$, is assumed to be less than that of the HR image, $N_{\boldsymbol{x}}$; i.e., $N_{\boldsymbol{y}}<N_{\boldsymbol{x}}$. We do this estimation using

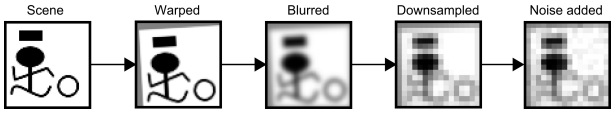


Fig. 1. An illustration of the image observation process

an SR technique whose resolution enhancement factor is $\alpha \equiv \sqrt{N_{\boldsymbol{x}}/N_{\boldsymbol{y}}}$ (> 1). Although we define the range of a pixel luminance value as infinite, we use -1 for black, $+1$ for white, and values between -1 and $+1$ for gradual gray.

The image observation process is modeled as shown in Fig. 1; the HR image $\boldsymbol{x}$ is geometrically warped, blurred, downsampled, and corrupted by noise $\boldsymbol{\epsilon}_l$ to form the observed LR image $\boldsymbol{y}_l$:

$$\boldsymbol{y}_l \equiv \boldsymbol{W}(\boldsymbol{\phi}_l)\boldsymbol{x} + \boldsymbol{\epsilon}_l, \tag{1}$$

or, more strictly,

$$p(\boldsymbol{Y}|\boldsymbol{x}, \beta, \boldsymbol{\Phi}) \equiv \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{y}_l; \boldsymbol{W}(\boldsymbol{\phi}_l)\boldsymbol{x}, \beta^{-1}\boldsymbol{I}). \tag{2}$$

The $\boldsymbol{\epsilon}_l \in \mathbb{R}^{N_{\boldsymbol{y}}}$ is AWGN with precision (inverse variance) β (> 0). Here, $\boldsymbol{W}(\boldsymbol{\phi}_l)$ is the $N_{\boldsymbol{y}} \times N_{\boldsymbol{x}}$ transformation matrix that is simultaneously used for warping, blurring, and downsampling. It is defined as

$$\boldsymbol{W}(\boldsymbol{\phi}_l)_{j,i} \equiv \frac{\mathcal{N}\left(\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right); 0, \gamma_l^{-1}\boldsymbol{I}\right)}{\sum_{i' \in \mathcal{I}} \mathcal{N}\left(\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_{i'}\right); 0, \gamma_l^{-1}\boldsymbol{I}\right)}, \tag{3}$$

$$\vec{\chi}\left(\theta, \vec{o}, \vec{\zeta}, \vec{\xi}\right) \equiv \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \left(\alpha\vec{\zeta} - \vec{o}\right) - \vec{\xi}, \tag{4}$$

where $\mathcal{I}$ represents the extent of the summation (explained in the next paragraph), and the vectors $\vec{\xi}_i$ and $\vec{\zeta}_j$ respectively denote the two-dimensional positions of the i-th pixel of the original HR image and the j-th pixel of the observed LR image. We define the center of each image as the origin and the size of each pixel is 1 by 1. For example, regarding an HR image with 40×40 pixels, each $\vec{\xi}$ represents $[-19.5, -19.5]^\top, [-18.5, -19.5]^\top, ..., [19.5, 19.5]^\top$. θ_l and $\vec{o}_l$ represent the warping parameters of the l-th LR image: the rotational motion parameter and translational motion parameter. The Gaussian distribution in (3) represents a Gaussian PSF that defines the blur, and γ_l (> 0) represents its precision parameter. In this paper, we assume γ_l also differs for each observed image. These transformation parameters are packed into $\boldsymbol{\phi}_l$, which is defined as

$$\boldsymbol{\Phi} \equiv \{\boldsymbol{\phi}_l\}_{l=1}^{L}, \ \boldsymbol{\phi}_l \equiv [\phi_{l,k}]_{k=1}^{4} \equiv [\theta_l, [\vec{o}_l]_h, [\vec{o}_l]_v, \gamma_l]^\top, \tag{5}$$

where subscripts h and v, respectively, denote horizontal and vertical positions on the image.

In previous works [3], [6], [8], the extent of $\mathcal{I}$ was defined as the extent of the HR image. According to this definition, however, the shape of the PSF is no longer Gaussian. For example, at the corner of the HR image, the shape is not omnidirectional but limited in a way such as that of a quadrant. In this paper, the extent of $\mathcal{I}$ is defined as infinite, and the luminance values outside the HR image are defined as 0 (middle gray). This normalization term faithfully represents the Gaussian PSF. We also found that this normalization term is exactly given by using the elliptic theta function ϑ_3, and we can rewrite $\boldsymbol{W}(\boldsymbol{\phi}_l)$ as

$$\boldsymbol{W}(\boldsymbol{\phi}_l)_{j,i} = \frac{\mathcal{N}\left(\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right); 0, \gamma_l^{-1}\boldsymbol{I}\right)}{\vartheta_3\left(\left[\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right)\right]_h, \mathrm{e}^{-\frac{2\pi^2}{\gamma_l}}\right)\vartheta_3\left(\left[\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right)\right]_v, \mathrm{e}^{-\frac{2\pi^2}{\gamma_l}}\right)}, \tag{6}$$

$$\vartheta_3(u, q) \equiv 1 + 2\sum_{n=1}^{\infty} q^{n^2} \cos 2n\pi u. \tag{7}$$

The elliptic theta function includes an infinite series, but it is easily determined numerically because the convergence is quite fast. In (6), the normalization term (the denominator of the right-hand side) seems to depend on i because $\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right)$ includes $\vec{\xi}_i$, but this is not true. Because the elliptic theta function is a periodic function with respect to the argument u with period 1, and $\vec{\chi}\left(\theta_l, \vec{o}_l, \vec{\zeta}_j, \vec{\xi}_i\right)$ can only take discrete values with step size 1 for the horizontal and vertical directions, the normalization term has the same value with respect to i.

C. HR Image Prior

Here, we introduce a "causal" Gaussian MRF prior for the HR image and additional latent variables. These latent variables are called the line process that controls the local correlation among pixel luminances. The introduction of the latent variables enables explicit expression of the possible discontinuity in the HR image. The line process, $\boldsymbol{\eta}$, consists of binary variables $\eta_{i,j} \in \{0, 1\}$ for all adjacent pixel pairs i and j. Its size equals $N_{\boldsymbol{\eta}} \equiv 2N_{\boldsymbol{x}} -$ [number of HR image's horizontal pixels] $-$ [number of HR image's vertical pixels]. We define the prior as

$$p(\boldsymbol{x}, \boldsymbol{\eta}|\lambda, \rho, \kappa) \equiv p(\boldsymbol{x}|\boldsymbol{\eta}, \rho, \kappa)p(\boldsymbol{\eta}|\lambda) \tag{8}$$

$$= \exp\left[-\lambda\sum_{i\sim j}(1-\eta_{i,j}) - \frac{\rho}{2}\sum_{i\sim j}\eta_{i,j}(x_i - x_j)^2 - \frac{\kappa}{2}\|\boldsymbol{x}\|_2^2 + \frac{1}{2}\ln\left|\frac{\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)}{2\pi}\right| + N_{\boldsymbol{\eta}}\ln\mathrm{logistic}(\lambda)\right], \tag{9}$$

where

$$p(\boldsymbol{\eta}|\lambda) \equiv \prod_{i\sim j}\mathrm{Bernoulli}\left(\eta_{i,j}; \mathrm{logistic}(\lambda)\right), \tag{10}$$

$$p(\boldsymbol{x}|\boldsymbol{\eta}, \rho, \kappa) \equiv \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)^{-1}), \tag{11}$$

$$\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)_{i,j} \equiv \begin{cases} \rho\sum_{k\sim i}\eta_{i,k} + \kappa, & i = j, \\ -\rho\eta_{i,j}, & i \sim j, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Here, the summation $\sum_{i\sim j}$ is taken over all pairs of adjacent pixels. The notation $i \sim j$ means that the i-th and j-th pixels are adjacent in the upward, downward, leftward, and rightward directions. The line process $\boldsymbol{\eta}$ switches the local characteristics

of the prior. It indicates whether two adjacent pixels take similar values or independent values. When $\eta_{i,j} = 1$, the i-th and the j-th pixels are strongly smoothed according to the quadratic penalty, whereas there is no smoothing when $\eta_{i,j} = 0$. The hyperparameter λ (> 0) is an edge penalty parameter that prevents $\eta_{i,j}$ from excessively taking edges. Note that λ is restricted to positive values because a negative λ leads to a reward rather than a penalty for taking edges. ρ (> 0) is a smoothness parameter that prevents the differences in adjacent pixel luminances from becoming large, and κ (> 0) is a contrast parameter that prevents $\boldsymbol{x}$ from taking an improperly large absolute value. On the other hand, in previous works [6], [8], κ is assumed to be 0, which results in an improper normalizing constant (see Discussion). $\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)$ is the $N_{\boldsymbol{x}} \times N_{\boldsymbol{x}}$ precision matrix of $\boldsymbol{x}$.

We have defined the introduced causal Gaussian MRF prior in the joint distribution form of $\boldsymbol{x}$ and $\boldsymbol{\eta}$, i.e., $p(\boldsymbol{\eta})p(\boldsymbol{x}|\boldsymbol{\eta})$. We call such a model "causal" because $\boldsymbol{\eta}$ seems to cause $\boldsymbol{x}$. The MRF model is defined as having the property

$$p(x_i|\boldsymbol{x}\backslash x_i, \boldsymbol{\eta}) = p(x_i|\boldsymbol{x}_{\mathcal{L}(i)}, \boldsymbol{\eta}_{i,\mathcal{L}(i)}) \tag{13}$$

in this case; i.e., the conditional distribution of a random variable, x_i, given all other variables, $\boldsymbol{x}\backslash x_i$ and $\boldsymbol{\eta}$, equals the conditional distribution of the random variable, x_i, given its "neighboring" variables, $\boldsymbol{x}_{\mathcal{L}(i)}$ and $\boldsymbol{\eta}_{i,\mathcal{L}(i)}$. If this conditional distribution is a Gaussian distribution, such an MRF is called a Gaussian MRF.

The "compound" MRF prior is usually defined in the form of the Gibbs distribution [18],

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{\eta}) \equiv \frac{\exp(-\tilde{H}(\boldsymbol{x}, \boldsymbol{\eta}))}{\sum_{\boldsymbol{\eta}} \int \exp(-\tilde{H}(\boldsymbol{x}, \boldsymbol{\eta}))\mathrm{d}\boldsymbol{x}}, \tag{14}$$

which is based on some microstate energy function, or equivalently, a Hamiltonian, such as

$$\begin{aligned}&\tilde{H}(\boldsymbol{x}, \boldsymbol{\eta})\\ &\equiv \lambda \sum_{i\sim j}(1-\eta_{i,j}) + \frac{\rho}{2}\sum_{i\sim j}\eta_{i,j}(x_i - x_j)^2 + \frac{\kappa}{2}\|\boldsymbol{x}\|_2^2.\end{aligned} \tag{15}$$

In addition to the property of (13), a compound MRF also has the property of

$$\tilde{p}(\eta_{i,j}|\boldsymbol{x}, \boldsymbol{\eta}\backslash\eta_{i,j}) = \tilde{p}(\eta_{i,j}|x_i, x_j), \tag{16}$$

whereas the introduced "causal" Gaussian MRF prior does not. Therefore, we do not call the introduced prior a "compound" MRF prior, even though (8) and (14) have similar forms. Furthermore, the introduced "causal" Gaussian MRF prior is a generative model, whereas the "compound" MRF is not. A generative model has the advantage of reducing the calculation cost (see Discussion).

D. Hyperparameter Prior

Generally, prior distributions should be non-informative unless we have explicit reasons because an informative prior leads to heuristics. Actually, we define the prior distributions for the hyperparameters of the HR image prior to be as non-informative as possible:

$$\begin{aligned}p(\lambda, \rho, \kappa, \beta) \equiv\ &\mathrm{Gamma}(\lambda; a_\lambda^{(0)}, b_\lambda^{(0)})\,\mathrm{Gamma}(\rho; a_\rho^{(0)}, b_\rho^{(0)})\\ &\times \mathrm{Gamma}(\kappa; a_\kappa^{(0)}, b_\kappa^{(0)})\,\mathrm{Gamma}(\beta; a_\beta^{(0)}, b_\beta^{(0)}),\end{aligned} \tag{17}$$

$$\begin{aligned}&a_\lambda^{(0)} \equiv 10^{-2}, b_\lambda^{(0)} \equiv 10^{-2},\ a_\rho^{(0)} \equiv 10^{-2}, b_\rho^{(0)} \equiv 10^{-2},\\ &a_\kappa^{(0)} \equiv 10^{-2}, b_\kappa^{(0)} \equiv 10^{-2},\ a_\beta^{(0)} \equiv 10^{-2}, b_\beta^{(0)} \equiv 10^{-2}.\end{aligned} \tag{18}$$

For a gamma distribution, the number of effective prior observations in the Bayesian framework is equal to two times parameter a. As shown in the Appendix, the number of observations for the hyperparameter λ is $N_{\boldsymbol{\eta}}$ in this SR. Also, that for ρ and κ is $N_{\boldsymbol{x}}$, and that for β is $LN_{\boldsymbol{y}}$. Therefore, the above settings – e.g., $2a_\lambda^{(0)} \ll N_{\boldsymbol{\eta}}$ – are considered sufficiently non-informative. Superscript (0) is added because we use these parameters as the initial values of variational Bayes later.

E. Registration Prior

For the registration parameters including the blurring parameter, we also define the corresponding prior as

$$p(\boldsymbol{\Phi}) \equiv \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{\phi}_l; \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(0)}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(0)}), \tag{19}$$

$$\boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(0)} \equiv [0, 0, 0, 12/\alpha^2],\ \boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(0)} \equiv \mathrm{diag}[10^{-3}, 10^0, 10^0, 10^{-3}]. \tag{20}$$

For the rotational motion parameter θ_l, the prior assumes 0 ± 1.81 degree ($\frac{180}{\pi}\sqrt{10^{-3}} \simeq 1.81$). This assumption is considered suitable for this SR task. Similarly, an assumption of 0 ± 1 pixels for translational motion parameters $[\vec{o}_l]_h$ and $[\vec{o}_l]_v$ is considered suitable. For blurring parameter γ_l, $\mu_{\gamma_l}^{(0)}$ is taken to be the value equivalent to the anti-aliasing of the scale factor α.

III. Objective Function and Estimator

A. Peak Signal-to-Noise Ratio (PSNR)

First, we confirm that the joint distribution of all random variables can now be explicitly given as

$$p(\boldsymbol{Y}, \boldsymbol{z}) = p(\boldsymbol{Y}|\boldsymbol{x}, \beta, \boldsymbol{\Phi})p(\boldsymbol{x}, \boldsymbol{\eta}|\lambda, \rho, \kappa)p(\lambda, \rho, \kappa, \beta)p(\boldsymbol{\Phi}), \tag{21}$$

$$\boldsymbol{z} \equiv [\boldsymbol{x}, \boldsymbol{\eta}, [\lambda, \rho, \kappa, \beta], \boldsymbol{\Phi}], \tag{22}$$

Once the joint distribution is obtained, we can derive all the marginal and conditional distributions; e.g., the posterior distribution $p(\boldsymbol{z}|\boldsymbol{Y})$ and joint distribution of the HR and LR images $p(\boldsymbol{Y}, \boldsymbol{x})$.

One of the most commonly used evaluation functions of the inferred image is the L2-norm (mean square error) -based PSNR. It is defined as

$$\mathrm{PSNR}(\hat{\boldsymbol{x}}; \boldsymbol{x}) \equiv 10\log_{10}\frac{2^2}{\frac{1}{N_{\boldsymbol{x}}}\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2}, \tag{23}$$

where $\hat{\boldsymbol{x}}$ is the estimator of the HR image and $\boldsymbol{x}$ is the true HR image. Since only LR images, $\boldsymbol{Y}$, are available for the estimator, we sometimes explicitly express it as a function

form, $\hat{\boldsymbol{x}}(\boldsymbol{Y})$. Now, our objective function (functional) to be maximized regarding the estimator is defined as

$$F(\hat{\boldsymbol{x}}(\boldsymbol{Y})) \equiv 10\log_{10}\frac{2^2}{\left\langle \frac{1}{N_{\boldsymbol{x}}}\|\hat{\boldsymbol{x}}(\boldsymbol{Y})-\boldsymbol{x}\|_2^2\right\rangle_{p(\boldsymbol{Y},\boldsymbol{x})}}. \tag{24}$$

This is because we prefer good estimator performance on average over various HR images and the corresponding LR images. Here, we assume that the occurrence rate of HR and LR images exactly coincides with the model we just introduced.

B. Posterior Mean (PM)

Using the above objective function, we can explicitly derive the best estimator of the HR image as the PM,

$$\underset{\hat{\boldsymbol{x}}(\boldsymbol{Y})}{\operatorname{argmax}} F(\hat{\boldsymbol{x}}(\boldsymbol{Y})) = \langle \boldsymbol{x}\rangle_{p(\boldsymbol{x}|\boldsymbol{Y})}. \tag{25}$$

Here, we used the well-known fact that the PM coincides with the minimum mean square error estimator in Bayesian framework. Note that $p(\boldsymbol{x}|\boldsymbol{Y})$ needs marginalization of all parameters other than $\boldsymbol{x}$ over $p(\boldsymbol{z}|\boldsymbol{Y})$. If the PM of the line process or other model parameters is necessary, it can also be determined in the same manner.

IV. Optimization Method

A. Variational Bayes

Though we could derive the optimal estimator, we cannot obtain the analytical solutions of the posterior distribution $p(\boldsymbol{z}|\boldsymbol{Y})$ and marginalized posterior distribution $p(\boldsymbol{x}|\boldsymbol{Y})$. Consequently, we have to rely on approximations. Here, we employ variational Bayes.

Variational Bayes [21] provides a trial distribution $q(\boldsymbol{z})$ that approximates the true posterior. We impose a factorization assumption on the trial distribution,

$$q(\boldsymbol{z}) \equiv q(\boldsymbol{x})q(\boldsymbol{\eta})q(\lambda,\rho,\kappa,\beta)q(\boldsymbol{\Phi}). \tag{26}$$

Note that, at this moment, the distribution family of each factorized distribution is not limited. We identify the optimal trial distribution that minimizes the KL divergence between the trial and the true distributions as the best approximation of the true distribution:

$$\hat{q}(\boldsymbol{z}) \equiv \underset{q(\boldsymbol{z})}{\operatorname{argmin}} D_{\mathrm{KL}}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{Y})). \tag{27}$$

Actually, the trial distribution that minimizes the KL divergence, not from $q(\boldsymbol{z})$ to $p(\boldsymbol{z}|\boldsymbol{Y})$ but from $p(\boldsymbol{z}|\boldsymbol{Y})$ to $q(\boldsymbol{z})$ coincides with the product of the exact marginal distributions as

$$\underset{q(\boldsymbol{z})}{\operatorname{argmin}} D_{\mathrm{KL}}(p(\boldsymbol{z}|\boldsymbol{Y})\|q(\boldsymbol{z})) = \prod_i p(z_i|\boldsymbol{Y}), \tag{28}$$

but this minimization is difficult to calculate.

Under the factorization assumption of the trial distribution and the extremal condition of the KL divergence, each optimal trial distribution should satisfy the self-consistent equations,

$$\hat{q}(z_i) \propto \exp\langle \ln p(\boldsymbol{z}|\boldsymbol{Y})\rangle_{\prod_{j\neq i}\hat{q}(z_j)}. \tag{29}$$

In the common style of variational Bayes [10], [23], this equation is solved by making repetitive updates,

$$q^{(0)}(z_i) \equiv p(z_i), \tag{30}$$

$$q^{(t+1)}(z_i) \propto \exp\langle \ln p(\boldsymbol{z}|\boldsymbol{Y})\rangle_{\prod_{j\neq i} q^{(t)}(z_j)}. \tag{31}$$

Each factorized trial distribution is supposed to converge to the optimal distribution. Sometimes, some $q^{(t+1)}(z_j)$s are used instead of $q^{(t)}(z_j)$s for the distribution on the right-hand side of (31). It depends on the hierarchical structure of the model. Similarly, some $q^{(0)}(z_i)$s may not be necessary.

B. Taylor Approximations

Although variational Bayes is a widely used general framework, its application is difficult in practice because it requires a conjugate prior. The prior distributions we have introduced are not conjugate priors. However, we have found that simple Taylor approximations make them conjugate and enable the analytical exact expectations in (31).

Here, to simplify the notation, we define the mean values of the latent variables $\boldsymbol{\eta}$, the hyper parameters $\lambda,\rho,\kappa,\beta$, and the registration parameters $\boldsymbol{\phi}_l$ over the trial distributions in the step number t of the updates of variational Bayes as $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)}$, $\mu_\lambda^{(t)}$, $\mu_\rho^{(t)}$, $\mu_\kappa^{(t)}$, $\mu_\beta^{(t)}$, $\boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)}$.

Specifically, we use first-order Taylor approximations for three non-linear terms. $\boldsymbol{W}(\boldsymbol{\phi}_l)$ is approximated around $\boldsymbol{\phi}_l = \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)}$,

$$\boldsymbol{W}(\boldsymbol{\phi}_l) \simeq \boldsymbol{W}_l^{(t)} + \sum_{k=1}^{4}[\boldsymbol{\phi}_l - \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)}]_k \boldsymbol{W}_{l,k}^{\prime(t)}, \tag{32}$$

where

$$\boldsymbol{W}_l^{(t)} \equiv \boldsymbol{W}(\boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)}), \tag{33}$$

$$\boldsymbol{W}_{l,k}^{\prime(t)} \equiv \left.\frac{\partial \boldsymbol{W}(\boldsymbol{\phi}_l)}{\partial \phi_{l,k}}\right|_{\boldsymbol{\phi}_l=\boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)}}. \tag{34}$$

Similarly, $\ln|\boldsymbol{A}(\boldsymbol{\eta},\rho,\kappa)|$ is approximated around $[\boldsymbol{\eta},\ln\rho,\ln\kappa] = [\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)},\ln\mu_\rho^{(t)},\ln\mu_\kappa^{(t)}]$,

$$\begin{aligned}\ln|\boldsymbol{A}(\boldsymbol{\eta},\rho,\kappa)| &\simeq \ln\left|\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)},\mu_\rho^{(t)},\mu_\kappa^{(t)})\right| \\ &+ \operatorname{tr}\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)},\mu_\rho^{(t)},\mu_\kappa^{(t)})^{-1}\Big[\mu_\rho^{(t)}\boldsymbol{A}(\boldsymbol{\eta}-\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)},1,0) \\ &+ (\ln\rho-\ln\mu_\rho^{(t)})\mu_\rho^{(t)}\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)},1,0) + (\ln\kappa-\ln\mu_\kappa^{(t)})\mu_\kappa^{(t)}\boldsymbol{I}\Big].\end{aligned} \tag{35}$$

We also use a similar approximation around $[\boldsymbol{\eta},\ln\rho,\ln\kappa] = [\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)},\ln\mu_\rho^{(t)},\ln\mu_\kappa^{(t)}]$. In addition, $\ln\operatorname{logistic}(\lambda)$ is approximated around $\ln\lambda = \ln\mu_\lambda^{(t)}$,

$$\begin{aligned}\ln\operatorname{logistic}(\lambda) &\simeq \ln\operatorname{logistic}(\mu_\lambda^{(t)}) \\ &+ (\ln\lambda-\ln\mu_\lambda^{(t)})\mu_\lambda^{(t)}\operatorname{logistic}(-\mu_\lambda^{(t)}).\end{aligned} \tag{36}$$

C. Update Equations

The trial distributions are obtained from (30)-(32), (35), and (36), as follows:

$$q^{(t)}(\boldsymbol{\eta}) = \prod_{i\sim j} \mathrm{Bernoulli}(\eta_{i,j}; \mu^{(t)}_{\eta_{i,j}}), \tag{37}$$

$$q^{(t)}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(t)}_{\boldsymbol{x}}, \boldsymbol{\Sigma}^{(t)}_{\boldsymbol{x}}), \tag{38}$$

$$\begin{aligned} q^{(t)}(\lambda,\rho,\kappa,\beta) = {} & \mathrm{Gamma}(\lambda; a^{(t)}_{\lambda}, b^{(t)}_{\lambda})\,\mathrm{Gamma}(\rho; a^{(t)}_{\rho}, b^{(t)}_{\rho}) \\ & \times \mathrm{Gamma}(\kappa; a^{(t)}_{\kappa}, b^{(t)}_{\kappa})\,\mathrm{Gamma}(\beta; a^{(t)}_{\beta}, b^{(t)}_{\beta}), \end{aligned} \tag{39}$$

$$q^{(t)}(\boldsymbol{\Phi}) = \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{\phi}_l; \boldsymbol{\mu}^{(t)}_{\boldsymbol{\phi}_l}, \boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\phi}_l}). \tag{40}$$

For (30) and (31), we update those distributions as follows. First, we compute $q^{(t+1)}(\boldsymbol{\eta})$ using $q^{(t)}(\boldsymbol{x}, \lambda, \rho, \kappa, \beta, \boldsymbol{\Phi})$. Second, we compute $q^{(t+1)}(\boldsymbol{x})$ using $q^{(t+1)}(\boldsymbol{\eta})q^{(t)}(\lambda, \rho, \kappa, \beta, \boldsymbol{\Phi})$. Finally, we compute $q^{(t+1)}(\lambda, \rho, \kappa, \beta)$ using $q^{(t+1)}(\boldsymbol{x}, \boldsymbol{\eta})q^{(t)}(\boldsymbol{\Phi})$ and $q^{(t+1)}(\boldsymbol{\Phi})$ using $q^{(t+1)}(\boldsymbol{x}, \boldsymbol{\eta})q^{(t)}(\lambda, \rho, \kappa, \beta)$. Here, we simply compute only the parameters of those distributions because we can compute the expectations in (31) analytically by using Taylor approximations in (32), (35), and (36). Specific update equations are described in the Appendix.

For the initial parameters of the trial distributions of $\boldsymbol{\eta}$ and $\boldsymbol{x}$, we use non-informative values,

$$\boldsymbol{\mu}^{(0)}_{\boldsymbol{\eta}} \equiv \mathbf{0},\ \boldsymbol{\mu}^{(0)}_{\boldsymbol{x}} \equiv \mathbf{0},\ \boldsymbol{\Sigma}^{(0)}_{\boldsymbol{x}} \equiv \mathbf{0}. \tag{41}$$

For the initial parameters for λ, ρ, κ, β and $\boldsymbol{\Phi}$, we use the same values as their prior's values.

We obtain the well-approximated PM of $\boldsymbol{x}$ as $\boldsymbol{\mu}^{(\infty)}_{\boldsymbol{x}}$. Realistically, instead of $\boldsymbol{\mu}^{(\infty)}_{\boldsymbol{x}}$, we use $\boldsymbol{\mu}^{(t+1)}_{\boldsymbol{x}}$ when the following convergence conditions hold for $\boldsymbol{\mu}^{(t+1)}_{\boldsymbol{x}}$ and each $\mu^{(t+1)}_{\phi_{l,k}}$,

$$\begin{aligned} & \frac{1}{N_{\boldsymbol{x}}} \|\boldsymbol{\mu}^{(t+1)}_{\boldsymbol{x}} - \boldsymbol{\mu}^{(t)}_{\boldsymbol{x}}\|_2^2 < 10^{-4}, \\ & \frac{1}{L}\sum_{l=1}^{L} \frac{(\mu^{(t+1)}_{\phi_{l,k}} - \mu^{(t)}_{\phi_{l,k}})^2}{[\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}]_k} < 10^{-4} \quad (k = 1, 2, 3, 4), \end{aligned} \tag{42}$$

where we defined $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}} \equiv [10^{-3}, 10^{0}, 10^{0}, 10^{-3}]$ as the scaling constant.

V. Experimental Results

The proposed method was evaluated using five gray-scale images with a size of 40×40 pixels, as shown in Fig. 2. From each image, $L = 10$ images with a size of 10×10 pixels were created by using (1), (2) with the settings of the parameters α, $\boldsymbol{\Phi}$, and β as the following. The resolution enhancement factor α was 4. The transformation parameter $\boldsymbol{\Phi}$ was randomly created according to the prior distribution in (19). The noise level parameter β was set for signal-to-noise ratios (SNR) of 20, 25, and 30 dB for each image. Samples of the created images are shown in Fig. 3.

Figure 4 shows the images estimated under SNR$=$ 30dB. The resolution of each image appeared to be better than the corresponding observed image in Fig. 3.

Table I lists the quantitative results compared to those from the methods of bilinear interpolation, Kanemura *et al.* [6], and Babacan *et al.* [10]. Note that we added a slight modification to these methods because they employ slightly different models. For example, the original method [10] assumes the blurring parameter γ is known, so we set γ as the mean value of the true distribution for this method. Also, we introduced a strong prior for λ in the Kanemura method [6] in contrast to the original method, because this parameter sometimes becomes negative. We evaluated the results with regard to the expectation and the standard deviation of the improvement in signal-to-noise ratio (ISNR) over 10 experiments on each image and for each SNR. ISNR is the relative PSNR defined as

$$\mathrm{ISNR} \equiv \mathrm{PSNR}(\hat{\boldsymbol{x}}; \boldsymbol{x}) - \mathrm{PSNR}(\tilde{\boldsymbol{x}}; \boldsymbol{x}), \tag{43}$$

where $\boldsymbol{x}$ is the true HR image, $\hat{\boldsymbol{x}}$ is the image estimated by the proposed method, and $\tilde{\boldsymbol{x}}$ is the image estimated by the compared method. A higher ISNR value means better improvement of the estimate against the estimate of the compared method. We see that the ISNRs of the proposed method were mostly higher than those of the other methods, except for the comparison with the Babacan's method in Pepper image.

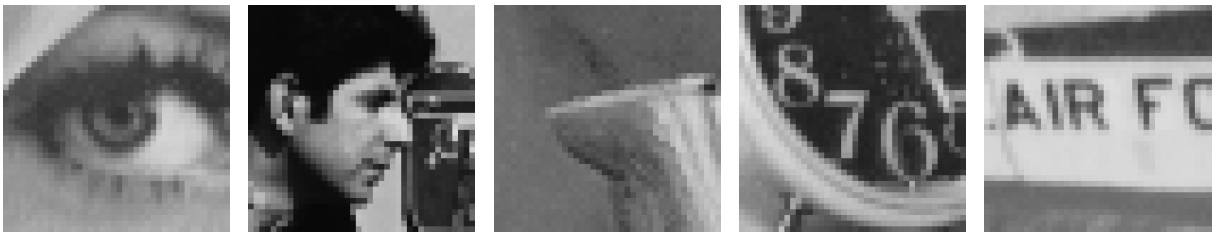
Table II lists the root mean square errors (RMSE) of our method and the other methods. To evaluate the estimated registration parameters, we took the RMSEs over 50 experiments (10 experiments $\times$ 5 images) for each noise level. Of course, a lower RMSE value means a better estimate. We see that the RMSEs of the proposed method were mostly higher than those of the other methods.

The calculation times of the proposed method was about 10 minutes on an Intel Core i7 2600 processor. The proposed method was a little slower than the method of Babacan *et al.* [10] and a little faster than the method of Kanemura *et al.* [6].

VI. Discussion

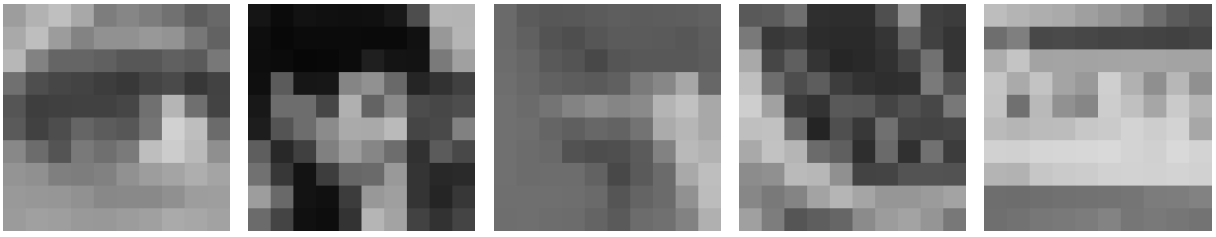
With regard to the observation model, we used a linear transformation and AWGN. The use of the linear transformation model is advantageous since an arbitrary transformation matrix $\boldsymbol{W}(\boldsymbol{\phi}_l)$ can be employed because of the Taylor approximation. The transformation matrix can be constructed by multiplying three matrices: the warping, blurring, and downsampling matrices [10]. A disadvantage of this is that sub-pixel errors might accumulate. We prefer matrix construction via a continuous function [3]. We improved the construction by introducing an elliptic theta function for the normalizing constant in (6). This normalizing constant provides fair pixel weights for both marginal and central areas of the HR image and faithfully represents the Gaussian PSF.

With regard to the HR image prior, we used a causal type of prior, which was first introduced by Kanemura *et al.* [6], [8]. The microstate energy function, or equivalently, the Hamiltonian, -based compound MRF prior of (14), offers the advantage of easy construction, but it usually has an exponential calculation cost, $\mathcal{O}(2^{N_{\boldsymbol{\eta}}})$, for the normalizing constant or, equivalently, the partition function, and this is an obstacle to direct calculation of the PM solution. The MAP solution has been used in work elsewhere because it is not affected by the normalizing constant. In contrast, the introduced causal type of prior of (8) has only a polynomial



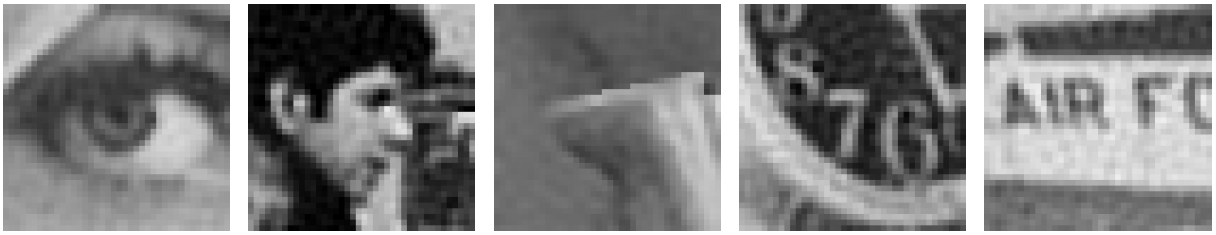
(a) Lena (b) Cameraman (c) Pepper (d) Clock (e) Text

Fig. 2. Five original images used in the experiments



(a) Lena (b) Cameraman (c) Pepper (d) Clock (e) Text

Fig. 3. Observed images when warped, blurred, downsampled by an enhancement factor of 4, and noised with SNR= 30dB AWGN



(a) Lena (b) Cameraman (c) Pepper (d) Clock (e) Text

Fig. 4. Images estimated from Fig. 3 observed images

calculation cost $\mathcal{O}(N_{\boldsymbol{x}}^3)$, which enables us to successfully apply the variational Bayes method to this problem.

With regard to the hyperparameter priors, we also improved the existing method. As the edge penalty parameter λ, Kanemura *et al.* [6] implicitly assumed $\lambda \in \mathbb{R}$, which leads to a negative λ and consequently results in an edge-strewn image. We assumed $\lambda > 0$ by setting its prior according to a gamma distribution, resulting in an appropriate inference. As the smoothness parameter ρ, they practically fixed the value of ρ with a strongly informative prior. We chose a non-informative prior for ρ. We show the box and whisker plot of the PM for each hyperparameter over 10 experiments on each image under SNR$=$ 30dB noise in Fig. 5. As can be seen, the inferred value of the PM of ρ showed wide variation, with an approximately 10-fold maximum-to-minimum ratio, depending on the original image. This result can be interpreted as meaning it is worth inferring ρ in each HR image. Furthermore, λ and κ respectively showed approximately 2-fold and 4-fold ranges of variation. Regarding the contrast parameter κ, they assumed $\kappa \equiv 0$, which leads to $|\boldsymbol{A}| = 0$, and this results in an improper normalizing constant. While we assume $\kappa > 0$, which leads to a proper normalizing constant, we can consequently take the term of $\ln |\boldsymbol{A}|$ into account in the update equations of the variational Bayes.

With regard to the prior distribution for the blurring parameter γ, we used a Gaussian distribution even though γ is a positive real number. This is because we selected a simpler expression. We tried using the prior of the gamma distribution as γ, but the improvement was small. One disadvantage of this model is that a non-informative setting for this prior may lead to a nonsense result where the inferred γ is negative. Moreover, we employed a somewhat informative prior for γ. This is because the blurring parameter γ and smoothness hyperparameter ρ are somewhat complementary. This means that simultaneous estimation of γ and ρ is difficult. Tipping *et al.* [3] and Kanemura *et al.* [6] fixed ρ, and Babacan *et al.* [9] fixed γ.

With regard to the estimator, we logically derived the optimal estimator PM from the objective function of the L2-norm-based PSNR. The widely used joint MAP estimator can be considered the optimal estimator for the all-or-none type objective function,

$$\underset{\hat{\boldsymbol{z}}}{\operatorname{argmax}} \left\langle \delta(\hat{\boldsymbol{z}} - \boldsymbol{z}) \right\rangle_{p(\boldsymbol{z}|\boldsymbol{Y})} = \underset{\boldsymbol{z}}{\operatorname{argmax}}\, p(\boldsymbol{z}|\boldsymbol{Y}), \qquad (44)$$

where δ is the Dirac delta or Kronecker delta function. Generally, this type of objective function is nonsensical for continuous variables because it is measure zero. If all the random variables in the posterior distribution are discrete, or if we can assume some smoothness of the posterior distribution,

TABLE I
PSNR OF THE PROPOSED METHOD (A HIGHER VALUE IS BETTER) AND ISNRS AGAINST THREE PREVIOUS METHODS (A HIGHER VALUE IS BETTER) FOR DIFFERENT IMAGES AND SNR LEVELS

Image	SNR [dB]	PSNR (proposed)	ISNR (vs bilinear)	ISNR (vs Kanemura)	ISNR (vs Babacan)
Lena	20	29.22 ± 0.33	$+5.35 \pm 0.33$	$+0.73 \pm 0.36$	-0.11 ± 0.08
	25	30.69 ± 0.25	$+6.78 \pm 0.30$	$+1.04 \pm 0.32$	$+0.11 \pm 0.11$
	30	32.13 ± 0.36	$+8.18 \pm 0.39$	$+1.60 \pm 0.37$	$+0.48 \pm 0.24$
Cameraman	20	21.74 ± 0.19	$+4.11 \pm 0.20$	$+1.06 \pm 0.36$	$+0.05 \pm 0.07$
	25	22.68 ± 0.21	$+5.00 \pm 0.23$	$+1.16 \pm 0.34$	-0.06 ± 0.08
	30	23.72 ± 0.38	$+6.04 \pm 0.39$	$+1.81 \pm 0.17$	$+0.03 \pm 0.06$
Pepper	20	29.71 ± 0.37	$+3.69 \pm 0.35$	$+0.11 \pm 0.17$	$+0.28 \pm 0.12$
	25	30.69 ± 0.28	$+4.57 \pm 0.27$	$+0.28 \pm 0.32$	$+0.06 \pm 0.16$
	30	31.23 ± 0.42	$+5.10 \pm 0.43$	$+0.81 \pm 0.75$	-0.34 ± 0.48
Clock	20	23.27 ± 0.18	$+5.36 \pm 0.19$	$+1.49 \pm 0.22$	$+0.11 \pm 0.10$
	25	24.29 ± 0.22	$+6.35 \pm 0.22$	$+1.77 \pm 0.26$	$+0.09 \pm 0.08$
	30	25.49 ± 0.50	$+7.53 \pm 0.51$	$+2.49 \pm 0.19$	$+0.32 \pm 0.13$
Text	20	24.67 ± 0.26	$+5.83 \pm 0.28$	$+1.57 \pm 0.20$	-0.10 ± 0.04
	25	25.87 ± 0.30	$+7.00 \pm 0.33$	$+1.98 \pm 0.33$	-0.03 ± 0.14
	30	27.26 ± 0.65	$+8.37 \pm 0.67$	$+3.03 \pm 0.42$	$+0.18 \pm 0.06$

a joint MAP solution will have meaning. Instead of the L2-norm-based objective function of PSNR, the L1-norm (mean absolute error) -based PSNR is sometimes employed. In such cases, the median of the posterior distribution is generally the optimal estimator. In the case of the marginalized ML, or equivalently, type-II ML or empirical Bayes, for example, the registration parameters and other hyperparameters are firstly inferred as:

$$[\hat{\lambda}, \hat{\rho}, \hat{\kappa}, \hat{\beta}, \hat{\boldsymbol{\Phi}}] \equiv \underset{\lambda,\rho,\kappa,\beta,\boldsymbol{\Phi}}{\operatorname{argmax}}\, p(\boldsymbol{Y}|\lambda, \rho, \kappa, \beta, \boldsymbol{\Phi}). \tag{45}$$

If these parameters have priors, such a method is called marginalized MAP. The HR image and sometimes the edge information are then inferred to as MAP,

$$\hat{\boldsymbol{x}} \equiv \underset{\boldsymbol{x}}{\operatorname{argmax}} \max_{\boldsymbol{\eta}} p(\boldsymbol{x}, \boldsymbol{\eta}|\boldsymbol{Y}, \hat{\lambda}, \hat{\rho}, \hat{\kappa}, \hat{\beta}, \hat{\boldsymbol{\Phi}}), \tag{46}$$

or PM. For such a two-step inference, it is difficult to calculate back the objective function.

With regard to the Taylor approximation for the transformation matrix $\boldsymbol{W}(\phi_l)$, we used the first-order approximation in (32) because it is more stable than the second-order approximation. This first-order approximation was proposed by Villena *et al.* [9]. The second-order approximation was proposed by Pickup *et al.* [5], and they obtained good results. We also tried the second-order approximation, but it sometimes made the algorithm unstable because it sometimes failed to produce a positive definite matrix for the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

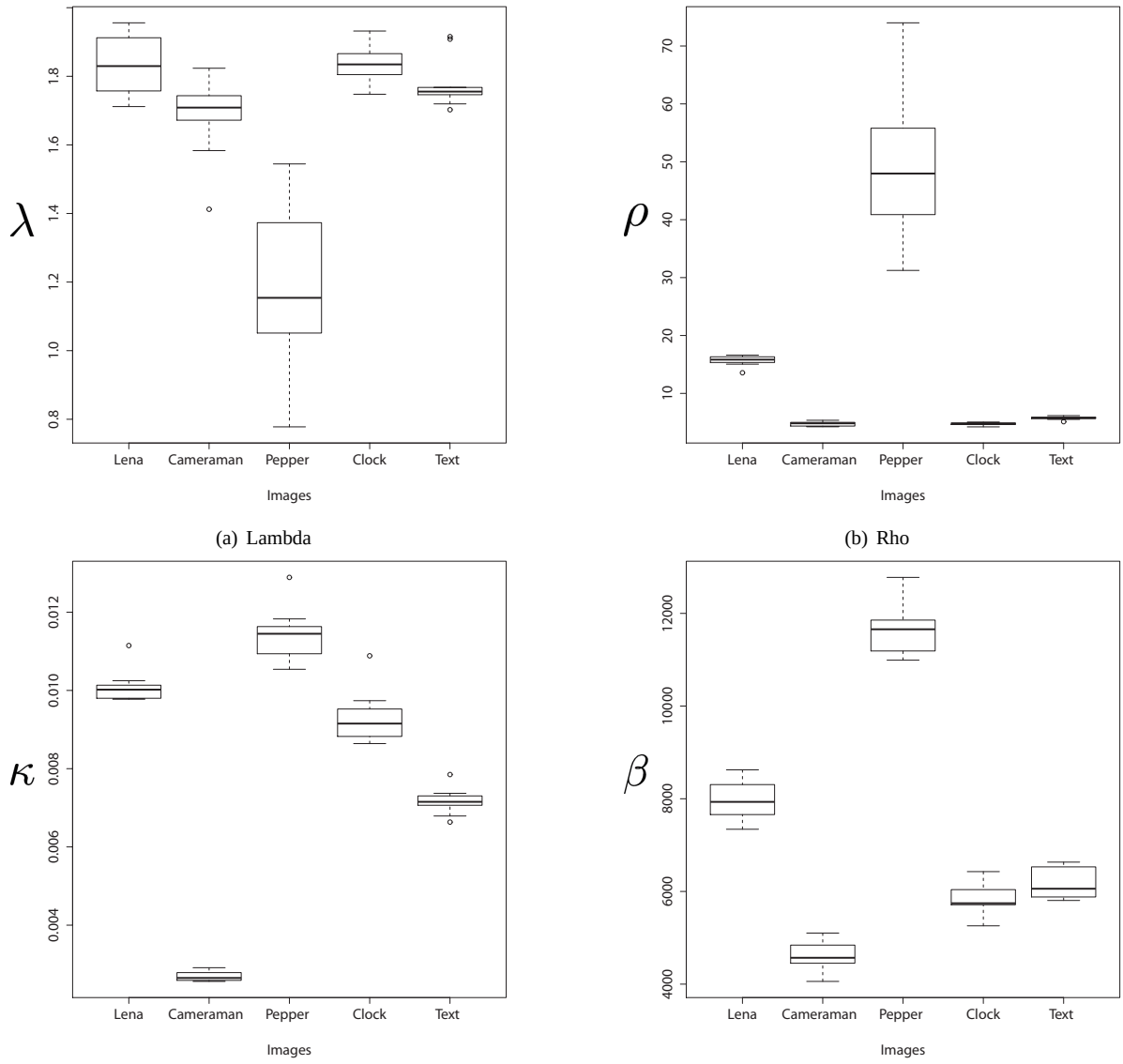
With regard to the Taylor approximation for $\ln|\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)|$ and $\ln \mathrm{logistic}(\lambda)$, we introduced the first-order approximation around $[\boldsymbol{\eta}, \ln\rho, \ln\kappa] = [\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)}, \ln\mu_{\rho}^{(t)}, \ln\mu_{\kappa}^{(t)}]$ and $\ln\lambda = \ln\mu_{\lambda}^{(t)}$, respectively, in (35) and (36). Note that the Taylor expansion not with respect to ρ, κ, λ, but with respect to $\ln\rho, \ln\kappa, \ln\lambda$ is our key idea to solve the conjugate prior problem. Indeed, we could successfully derive the terms originating from $\ln|\boldsymbol{A}|$ in update equations ((52), (60), and (62) in Appendix). Kanemura *et al.* [6], [8] ignored the term of $\ln|\boldsymbol{A}|$ because of the high calculation cost, and this would result in less accurate inference. As for $\boldsymbol{\eta}$, we implicitly assumed that $\boldsymbol{\eta}$ is not a binary vector but a continuous vector and did the differentiation. This assumption is based on (12). If we make another assumption – i.e., replacement of $\eta_{i,j}$ with $\eta_{i,j}^2$ in (12) – (12) has the same meaning, but the result of the Taylor approximation will differ from the current form.

With regard to the experimental results, the proposed method outperforms the other methods in terms of the ISNR for most images and noise levels. Moreover, its estimation of the registration parameters was more accurate than the other methods were for most conditions. Therefore, we conclude the proposed method is on the whole superior to the other methods. Compared with bilinear interpolation and Kanemura's method, the superiority of the proposed method was clear. Compared with the Babacan's method, the superiority of the proposed method was rather slight. Especially, in the case of the Pepper image in 30 dB noise, the porposed method was worse than the Babacan's method. This inferiority is considered to be caused by unstable estimation of γ and ρ, where Babacan's method fixed the value of γ to the true expected value in our implementation. Intuitively, the Pepper image is smoother than the other images and has fewer edges. Therefore, this feature is considered to be less preferable for complementary parameters of γ and ρ.

With regard to the calculation cost, the proposed algorithm requires $\mathcal{O}(N_{\boldsymbol{x}}^3)$. This calculation cost order is given by two matrix inversions: $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{(t+1)}$ in (55) and $\boldsymbol{A}$ in (52) and (62) (see Appendix). We found that a simple approximation such as considering all the off-diagonal elements to be zero reduces the calculation time but obviously degrades accuracy. We hope to solve this problem in our future work.

VII. CONCLUSION

In this paper, we proposed a Bayesian image super-resolution (SR) method with a causal Gaussian Markov random field (MRF) prior. We improved existing models with respect to three points: 1) the combined transformation model through a preferable normalization term using the elliptic theta function, 2) the causal Gaussian MRF model through introduction of a contrast parameter κ, which provides an effective normalizing constant including $\ln|\boldsymbol{A}|$, and 3) the hyperparameter prior model through application of a gamma



(a) Lambda (b) Rho

(c) Kappa (d) Beta

Fig. 5. Box and whisker plot of the PM for each hyperparameter, λ, ρ, κ, and β, and image under SNR$=$ 30dB noise

TABLE II
RMSEs OF REGISTRATION PARAMETERS (A LOWER VALUE IS BETTER) FOR DIFFERENT SNR LEVELS

parameter	SNR [dB]	RMSE (proposed)	RMSE (Kanemura)	RMSE (Babacan)
θ	20	0.006	0.006	0.006
	25	0.004	0.004	0.004
	30	0.002	0.003	0.003
$[\vec{o}]_h$	20	0.094	0.095	0.094
	25	0.054	0.059	0.056
	30	0.041	0.060	0.046
$[\vec{o}]_v$	20	0.074	0.073	0.076
	25	0.044	0.052	0.047
	30	0.037	0.044	0.036
γ	20	0.031	0.033	—
	25	0.025	0.030	—
	30	0.028	0.028	—

distribution for the edge penalty parameter λ, which prevents an unfavorable edge-strewn image. We then logically derived the optimal estimator, that is, not the joint maximum a posteriori (MAP) or marginalized maximum likelihood (ML) but the posterior mean (PM), from the objective function of the L2-norm (mean square error) -based peak signal-to-noise ratio (PSNR). The estimator is numerically determined by using variational Bayes. We solved the conjugate prior problem in variational Bayes by introducing three Taylor approximations. Other than these approximations, we did not use any approximations such as ignoring the term $\ln|\boldsymbol{A}|$. Experimental results showed that the proposed method is mostly superior to existing methods in accuracy.

APPENDIX

Here, we show the details of the variational Bayes' update equations in Section IV-C.

The mean values of the hyperparameters $\lambda, \rho, \kappa, \beta$ over the trial distributions $q^{(t)}(\lambda, \rho, \kappa, \beta)$ are given by

$$\mu_\lambda^{(t)} = \frac{a_\lambda^{(t)}}{b_\lambda^{(t)}},\ \mu_\rho^{(t)} = \frac{a_\rho^{(t)}}{b_\rho^{(t)}},\ \mu_\kappa^{(t)} = \frac{a_\kappa^{(t)}}{b_\kappa^{(t)}},\ \mu_\beta^{(t)} = \frac{a_\beta^{(t)}}{b_\beta^{(t)}}. \tag{47}$$

The update equation of $\boldsymbol{\eta}$ is given as

$$\begin{aligned} q^{(t+1)}(\boldsymbol{\eta}) &\propto \exp \left\langle \ln p(\boldsymbol{z}|\boldsymbol{Y}) \right\rangle_{q^{(t)}(\boldsymbol{x},\lambda,\rho,\kappa,\beta,\boldsymbol{\Phi})} \\ &\propto \exp \left(\sum_{i\sim j} \left\{ c_\lambda^{(t)} - \frac{\mu_\rho^{(t)}}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t)} \boldsymbol{M}_{i,j} \right\} \eta_{i,j} \right. \\ &\quad \left. + \frac{1}{2} \left\langle \ln |\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)| \right\rangle_{q^{(t)}(\rho,\kappa)} \right), \end{aligned} \tag{48}$$

where

$$\boldsymbol{C}_{\boldsymbol{x}}^{(t)} \equiv \boldsymbol{\mu}_{\boldsymbol{x}}^{(t)} [\boldsymbol{\mu}_{\boldsymbol{x}}^{(t)}]^\top + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{(t)}, \tag{49}$$

$$[\boldsymbol{M}_{i,j}]_{k,l} \equiv \begin{cases} +1, & (k,l) = (i,i) \text{ or } (j,j), \\ -1, & (k,l) = (i,j) \text{ or } (j,i), \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

Using the Taylor approximation of (35), we obtain the distribution of (37) at step $t+1$ with the parameter

$$\mu_{\eta_{i,j}}^{(t+1)} = \operatorname{logistic} \left(\mu_\lambda^{(t)} + \frac{1}{2} \mu_\rho^{(t)} C_{\eta_{i,j}}^{(t)} \right), \tag{51}$$

where

$$C_{\eta_{i,j}}^{(t)} \equiv \operatorname{tr} \left[\left(\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t)}, \mu_\rho^{(t)}, \mu_\kappa^{(t)})^{-1} - \boldsymbol{C}_{\boldsymbol{x}}^{(t)} \right) \boldsymbol{M}_{i,j} \right]. \tag{52}$$

The update equation of $\boldsymbol{x}$ is given as

$$\begin{aligned} q^{(t+1)}(\boldsymbol{x}) &\propto \exp \left\langle \ln p(\boldsymbol{z}|\boldsymbol{Y}) \right\rangle_{q^{(t+1)}(\boldsymbol{\eta}) q^{(t)}(\lambda,\rho,\kappa,\beta,\boldsymbol{\Phi})} \\ &\propto \exp \left(-\frac{1}{2} \left\{ \boldsymbol{x}^\top \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, \mu_\rho^{(t)}, \mu_\kappa^{(t)}) \boldsymbol{x} \right. \right. \\ &\quad \left. \left. + \mu_\beta^{(t)} \sum_{l=1}^{L} \left\langle \|\boldsymbol{W}(\boldsymbol{\phi}_l)\boldsymbol{x} - \boldsymbol{y}_l\|_2^2 \right\rangle_{q^{(t)}(\boldsymbol{\phi}_l)} \right\} \right). \end{aligned} \tag{53}$$

It becomes a Gaussian distribution. Using the Taylor approximation (32), we obtain the distribution of (38) at step $t+1$ with the parameters

$$\boldsymbol{\mu}_{\boldsymbol{x}}^{(t+1)} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{(t+1)} \left[\mu_\beta^{(t)} \sum_{l=1}^{L} \boldsymbol{y}_l^\top \boldsymbol{W}_l^{(t)} \right]^\top, \tag{54}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}^{(t+1)} = \left[\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, \mu_\rho^{(t)}, \mu_\kappa^{(t)}) + \mu_\beta^{(t)} \sum_{l=1}^{L} \boldsymbol{C}_{\boldsymbol{W}_l}^{\prime(t)} \right]^{-1}, \tag{55}$$

where

$$\boldsymbol{C}_{\boldsymbol{W}_l}^{\prime(t)} \equiv [\boldsymbol{W}_l^{(t)}]^\top \boldsymbol{W}_l^{(t)} + \sum_{k,k'} [\boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(t)}]_{k,k'} [\boldsymbol{W}_{l,k}^{\prime(t)}]^\top \boldsymbol{W}_{l,k'}^{\prime(t)}. \tag{56}$$

The update equation of $\lambda, \rho, \kappa, \beta$ is given as

$$\begin{aligned} q^{(t+1)}(\lambda, \rho, \kappa, \beta) &\propto \exp \left\langle \ln p(\boldsymbol{z}|\boldsymbol{Y}) \right\rangle_{q^{(t+1)}(\boldsymbol{x},\boldsymbol{\eta}) q^{(t)}(\boldsymbol{\Phi})} \\ &\propto \lambda^{a_\lambda^{(0)}-1} \rho^{a_\rho^{(0)}-1} \kappa^{a_\kappa^{(0)}-1} \beta^{a_\beta^{(0)}+\frac{1}{2}LN_{\boldsymbol{y}}} \exp \left(\right. \\ &\quad - \left\{ b_\lambda^{(0)} + \sum_{i\sim j} (1 - \mu_{\eta_{i,j}}^{(t+1)}) \right\} \lambda \\ &\quad - \left\{ b_\rho^{(0)} + \frac{1}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, 1, 0) \right\} \rho \\ &\quad - \left\{ b_\kappa^{(0)} + \frac{1}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \right\} \kappa \\ &\quad - \left\{ b_\beta^{(0)} + \frac{1}{2} \sum_{l=1}^{L} \left\langle \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \boldsymbol{W}(\boldsymbol{\phi}_l)^\top \boldsymbol{W}(\boldsymbol{\phi}_l) \right. \right. \\ &\qquad \left. \left. - 2\boldsymbol{y}_l^\top \boldsymbol{W}(\boldsymbol{\phi}_l) \boldsymbol{\mu}_{\boldsymbol{x}}^{(t+1)} + \boldsymbol{y}_l^\top \boldsymbol{y}_l \right\rangle_{q^{(t)}(\boldsymbol{\phi}_l)} \right\} \beta \\ &\quad \left. + \frac{1}{2} \left\langle \ln |\boldsymbol{A}(\boldsymbol{\eta}, \rho, \kappa)| \right\rangle_{q^{(t+1)}(\boldsymbol{\eta})} + N_{\boldsymbol{\eta}} \ln \operatorname{logistic}(\lambda) \right). \end{aligned} \tag{57}$$

Using Taylor approximations (32), (35), and (36), we obtain the distribution of (39) at step $t+1$ with parameters

$$a_\lambda^{(t+1)} = a_\lambda^{(0)} + N_{\boldsymbol{\eta}} \mu_\lambda^{(t)} \operatorname{logistic}(-\mu_\lambda^{(t)}), \tag{58}$$

$$b_\lambda^{(t+1)} = b_\lambda^{(0)} + \sum_{i\sim j} (1 - \mu_{\eta_{i,j}}^{(t+1)}), \tag{59}$$

$$a_\rho^{(t+1)} = a_\rho^{(0)} + \frac{\mu_\rho^{(t)}}{2} \operatorname{tr} \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, \mu_\rho^{(t)}, \mu_\kappa^{(t)})^{-1} \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, 1, 0) \tag{60}$$

$$b_\rho^{(t+1)} = b_\rho^{(0)} + \frac{1}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, 1, 0), \tag{61}$$

$$a_\kappa^{(t+1)} = a_\kappa^{(0)} + \frac{\mu_\kappa^{(t)}}{2} \operatorname{tr} \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{(t+1)}, \mu_\rho^{(t)}, \mu_\kappa^{(t)})^{-1} \tag{62}$$

$$b_\kappa^{(t+1)} = b_\kappa^{(0)} + \frac{1}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)}, \tag{63}$$

$$a_\beta^{(t+1)} = a_\beta^{(0)} + \frac{1}{2} LN_{\boldsymbol{y}}, \tag{64}$$

$$b_\beta^{(t+1)} = b_\beta^{(0)} + \frac{1}{2} \sum_{l=1}^{L} \left(\operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \boldsymbol{C}_{\boldsymbol{W}_l}^{\prime(t)} - 2\boldsymbol{y}_l^\top \boldsymbol{W}_l^{(t)} \boldsymbol{\mu}_{\boldsymbol{x}}^{(t+1)} + \boldsymbol{y}_l^\top \boldsymbol{y}_l \right). \tag{65}$$

The update equation of $\boldsymbol{\Phi}$ is given as

$$\begin{aligned} q^{(t+1)}(\boldsymbol{\Phi}) &\propto \exp \left\langle \ln p(\boldsymbol{z}|\boldsymbol{Y}) \right\rangle_{q^{(t+1)}(\boldsymbol{x},\boldsymbol{\eta}) q^{(t)}(\lambda,\rho,\kappa,\beta)} \\ &\propto \exp \left(-\frac{1}{2} \sum_{l=1}^{L} \left\{ [\boldsymbol{\phi}_l - \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(0)}]^\top [\boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(0)}]^{-1} [\boldsymbol{\phi}_l - \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(0)}] \right. \right. \\ &\quad \left. \left. + \mu_\beta^{(t)} \left\{ \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \boldsymbol{W}(\boldsymbol{\phi}_l)^\top \boldsymbol{W}(\boldsymbol{\phi}_l) - 2\boldsymbol{y}_l^\top \boldsymbol{W}(\boldsymbol{\phi}_l) \boldsymbol{\mu}_{\boldsymbol{x}}^{(t+1)} \right\} \right\} \right). \end{aligned} \tag{66}$$

Using the Taylor approximation (32), we obtain the distribu-

tion of (40) at step $t+1$ with parameters

$$\boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t+1)} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(t+1)} \left[[\boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(0)}]^{-1} \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(0)} + \mu_\beta^{(t)} [\boldsymbol{C}_{\boldsymbol{\phi}_l}^{\prime\prime(t+1)} \boldsymbol{\mu}_{\boldsymbol{\phi}_l}^{(t)} - \boldsymbol{C}_{\boldsymbol{\phi}_l}^{\prime(t+1)}] \right], \tag{67}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(t+1)} = \left[[\boldsymbol{\Sigma}_{\boldsymbol{\phi}_l}^{(0)}]^{-1} + \mu_\beta^{(t)} \boldsymbol{C}_{\boldsymbol{\phi}_l}^{\prime\prime(t+1)} \right]^{-1}, \tag{68}$$

where

$$[\boldsymbol{C}_{\boldsymbol{\phi}_l}^{\prime(t+1)}]_k \equiv \frac{1}{2} \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} \left[[\boldsymbol{W}_l^{(t)}]^\top \boldsymbol{W}_{l,k}^{\prime(t)} + [\boldsymbol{W}_{l,k}^{\prime(t)}]^\top \boldsymbol{W}_l^{(t)} \right] - \boldsymbol{y}_l^\top \boldsymbol{W}_{l,k}^{\prime(t)} \boldsymbol{\mu}_{\boldsymbol{x}}^{(t+1)}, \tag{69}$$

$$[\boldsymbol{C}_{\boldsymbol{\phi}_l}^{\prime\prime(t+1)}]_{k,k'} \equiv \operatorname{tr} \boldsymbol{C}_{\boldsymbol{x}}^{(t+1)} [\boldsymbol{W}_{l,k}^{\prime(t)}]^\top \boldsymbol{W}_{l,k'}^{\prime(t)}. \tag{70}$$

REFERENCES

[1] R. Tsai and T. Huang, "Multiframe image restoration and registration," in *Advances in computer vision and image processing*, vol. 1, no. 2, JAI Press Inc., pp. 317–339, Greenwich, CT, 1984.

[2] R. Hardie, K. Barnard, and E. Armstrong, "Joint Map registration and high resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, 1997.

[3] M. E. Tipping, and C. M. Bishop, "Bayesian image super-resolution," in *Advances in NIPS 15*, MIT Press, pp. 1279–1286, 2003.

[4] R. Molina, J. Mateos, A. K. Katsaggelos, and M. Vega, "Bayesian multichannel image restoration using compound Gauss-Markov random fields," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1642–1654, 2003.

[5] L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman, "Bayesian Image Super-Resolution, continued," in *Advances in NIPS 19*, MIT Press, 2007.

[6] A. Kanemura, S. Maeda, and S. Ishii, "Hyperparameter Estimation in Bayesian Image Superresolution with a Compound Markov Random Field Prior," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 181–186, 2007.

[7] P. Vandewalle, L. Sbaiz, J. Vandewalle, and M. Vetterli, "Super-resolution from unregistered and totally aliased signals using subspace methods," *IEEE Trans. Signal Processing*, vol. 55, No. 7, Part 2, pp. 3687–3703, 2007.

[8] A. Kanemura, S. Maeda, and S. Ishii, "Superresolution with compound Markov random fields via the variational EM algorithm," *Neural Networks*, vol. 22, pp. 1025–1034, 2009.

[9] S. Villena, M. Vega, S.D. Babacan, R. Molina, and A. K. Katsaggelos, "Image Prior Combination in Super-resolution Image Registration & Reconstruction," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 355–360, 2010.

[10] S. D. Babacan, R. Molina, A. K. Katsaggelos, "Variational Bayesian Super Resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[11] S. Borman and R L. Stevenson, "Spatial resolution enhancement of low-resolution image sequences a comprehensive review with directions for future research," Department of Electrical Engineering, University of Notre Dame, Tech. Rep., 1998.

[12] S. C. Park, M. K. Park, M. G. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine*, IEEE, vol. 20, no.3, pp. 21–36, 2003.

[13] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *Int. J. Imag. Syst. Technol.*, vol. 14, no. 2, pp. 47–57, 2004.

[14] M. Ng, T. Chan, M. G. Kang, and P. Milanfar, "Special Issue on Super-resolution Imaging: Analysis, Algorithms, and Applications," *EURASIP Journal on Applied Signal Processing*, 2006.

[15] A.K. Katsaggelos, R. Molina, and J. Mateos, "Super Resolution of Images and Video", Synthesis Lectures on Image, Video, and Multimedia Processing, Morgan & Claypool, 2007.

[16] P. Milanfar, Ed., "Super-Resolution Imaging," CRC Press, 2010.

[17] R. Molina, M. Vega, J. Abad, and A. K. Katsaggelos, "Parameter Estimation in Bayesian High-Resolution Image Reconstruction With Multisensors," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1655–1667, 2003.

[18] S. Geman, and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI–6, no. 6, pp. 721–741, 1984.

[19] R. Chellappa, and A.K. Jain, Eds., "Markov random fields: Theory and application", Academic Press, Boston 1993.

[20] F. -C. Jeng, and J. W. Woods, "Compound Gauss-Markov random fields for image estimation", *IEEE Transactions on Signal Processing*, vol. 39, no. 3, pp. 683–697, 1991.

[21] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. UAI*, San Francisco, CA, pp. 21–30, Morgan Kaufmann, 1999.

[22] F. -C. Jeng, and J. W. Woods, "Simulated annealing in compound Gaussian random fields," *IEEE Transactions on Information Theory*, vol. 36, no. 1, pp. 94–107, 1990.

[23] C. M. Bishop, D. Spiegelhalter, J. Winn "VIBES: A Variational Inference Engine for Bayesian Networks," in *Advances in NIPS 15*, MIT Press, pp. 777–784, 2003.